View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by LUISSearch

Bayesian Likelihoods for Moment Condition Models

Giuseppe Ragusa
Department of Economics
University of California, Irvine

January 10, 2007

Abstract

Bayesian inference in moment condition models is difficult to implement. For these models, a posterior distribution cannot be calculated because the likelihood function has not been fully specified. In this paper, we obtain a class of likelihoods by formal Bayesian calculations that take into account the semiparametric nature of the problem. The likelihoods are derived by integrating out the nuisance parameters with respect to a maximum entropy tilted prior on the space of distribution. The result is a unification that uncovers a mapping between priors and likelihood functions. We show that there exist priors such that the likelihoods are closely connected to Generalized Empirical Likelihood (GEL) methods.

Keywords: Moment condition, GMM, GEL, Likelihood functions, Approximate Bayesian inference.

1 Introduction

A typical Bayesian model is composed of a parametric likelihood function and a prior distribution on the parameters. Bayes' theorem then actualizes the information on the parameters by using the data. In practical situations, however, one may be unwilling or unable to specify a fully parametric likelihood. For example, a researcher's interest may be in an economic theory that makes weak structural predictions or robustness concerns may discourage a committment to a particular parametric form.

This paper investigates the viability of conducting Bayesian inference when the only information linking the parameters and the data is in the form of moment restrictions. We propose a way of obtaining likelihoods in moment condition models by using formal Bayesian calculations. The basic idea is to simplify the problem by making nonparametric assumptions on the set of distributions supported by the moment condition. An initial prior is attached to the space of distributions and is then tilted to make it satisfy the moment condition. The tilted prior minimizes the Kullback-Leibler divergence between the initial prior and all the distributions supported by the model. The tilted distribution is then used to integrate out the nuisance parameters, which in this context are multinomial weights.

We show that there is a mapping between the initial priors and the resulting likelihoods. There exist prior distributions such that the likelihoods are related to Generalized Empirical Likelihood (GEL) methods (Newey and Smith, 2004). GEL is a generalization of the empirical likelihood (EL) (Qin and Lawless, 1994; Imbens, 1997) and exponential tilting (ET) (Kitamura and Stutzer, 1997). A feature of GEL is that it provides semiparametric efficient estimators of the cumulative distribution function (c.d.f.) of the data under the set of moment restrictions (Brown and Newey, 2002). It is fitting that there exist priors such that resulting likelihoods from our approach are functionally related to these estimators. These likelihoods are the product of the GEL weights that define the semiparametric efficient estimator of the c.d.f.

There is recent literature on Bayesian inference in semiparametric models that is related to our work. Lazar (2003) considers using the product of the EL weights as a likelihood in the posterior distribution. Schennach (2005) obtains an integrated likelihood by an asymptotic procedure in which the nuisance parameters grow to infinity. Schennach's likelihood is related to a member of the GEL class: it is the product of the ET weights. The method dis-

cussed in this paper provides a probabilistic justification for the approach in Lazar (2003) and it extends Schennach (2005) by showing that Bayesian likelihoods can also be constructed from other members of the GEL class.

The approach of Chamberlain and Imbens (2003) is also related to our work, which includes extending the Bayesian bootstrap (BB) of Rubin (1981) to moment condition models. Our approach and the approach of Chamberlain and Imbens (2003) share the same set of nonparametric assumptions. In particular, in both cases the model is restricted to discrete random variables whose support is assumed to be fully observed in the data. However, while the Bayesian bootstrap obtains samples from the posterior distribution by assuming an improper Dirichlet prior on the parameters of the multinomial distribution, we obtain a posterior by integrating out the parameters of the multinomial with respect to a prior that carries information contained in the model.

As noted by Sims (2002), if moment condition models are going to be used in real decision making, classical confidence bands for parameters are going to be interpreted as posterior probability credible regions. It is therefore important to know under which assumptions the use of semiparametric likelihoods gives valid Bayesian inference. This paper sheds some light on this issue by showing explicitly what class of priors on the nuisance parameters supports the use of GEL-based likelihoods in Bayesian inference.

The effect of the nonparametric assumptions on the resulting inference could be a concern. This issue is addressed by investigating whether credible regions based on posteriors that use semiparametric likelihoods have correct coverage. We explore coverage properties in two simple settings. The first is concerned with the estimation of quantiles of a continuous distribution. This example shows that when the moment condition is bounded the choice of the prior for the nuisance parameters is immaterial: all the semiparametric likelihoods lead to the same Bayesian inference. The second example is an overidentified location problem. In this case there are differences both in the inference and in the coverage of the likelihoods. We find that semiparametric likelihoods give valid Bayesian inference for reasonable sample sizes. One exception is the case of the likelihood obtained by eliciting a normal initial prior on the multinomial probabilities. In this case, the likelihood corresponds to the product of the weights of the continuous updating (CUE). This suggests that the elicitation of the initial prior for the nuisance parameters may have a larger effect on the validity of the inference

than the nonparametric assumptions.

This paper can also be interpreted as an attempt to reconcile classical and Bayesian estimation of parameters specified through a moment condition. Chernozhukov and Hong (2003) develop an asymptotic theory for estimators defined as means and quantiles of quasi-posteriors based on statistical criterion functions. Their analysis can be extended to the likelihoods obtained here. The resulting inference could then be interpreted from a Bayesian or a classical perspective depending on the objective of the analysis. Ragusa (2006a) analyzes the frequentist properties of the semiparametric likelihoods and studies their performance both in Monte Carlo experiments and in real applications.

The remainder of the paper proceeds as follows. Section 2 formally defines the moment condition models, introduces the basic notation, and briefly reviews the literature. Section 3 develops the Bayesian calculations and derive the semiparametric likelihoods. Section 4 establishes the relationship between the likelihoods of Section 3 and GEL methods. Section 5 explores the Bayesian coverage validity of the nonparametric likelihoods. Finally, Section 6 concludes and points to future work.

2 Model and motivations

The model we consider is for iid observations where there is a countable number of moment restrictions. To describe the model, let $x_i (i = 1, \ldots, n)$ be i.i.d. observations on a random vector x. Also, let β be a $p \times 1$ parameter vector and $g(x, \beta) = (g_1(x, \beta), \ldots, g_m(x, \beta))'$ an $m \times 1$ vector of functions of x and the parameter vector, where $m \geq p$. At the true parameter vector, β_0, the model satisfies the moment condition

$$\int g(x, \beta_0) dF(x) = 0 \tag{1}$$

where $F(x)$ denotes the distribution of x. The moment equation (1) is implied by a conditional restriction and in general most models considered in econometrics fit this framework.

Classical inference proceeds by applying the efficient GMM procedure to obtain consistent estimators of β and then constructing statistics based on the large sample distribution of the estimator. To describe this procedure, let $g_i(\beta) = g(x_i, \beta)$, $\bar{g}(\beta) = \sum_{i=1}^{n} g_i(x_i, \beta)/n$

and $\bar{\Omega}(\beta) = \sum_{i=1}^{n} g(x_i, \beta) g(x_i, \theta)'/n$. The efficient two-step GMM estimator is given by

$$\hat{\beta}^{GMM} = \arg\min_{\beta \in \mathcal{B}} J(\beta; \bar{\beta}), \quad J(\beta; \bar{\beta}) = n \bar{g}(\beta)' \bar{W} \bar{g}(\beta), \quad \bar{W} = \bar{\Omega}(\beta)^{-1},$$

where $\bar{\beta}$ is a preliminary consistent estimate of β and $\mathcal{B}$ is a compact set of parameter values. In a wide array of settings, $\sqrt{n}(\hat{\beta}^{GMM} - \beta_0) = O_p(1)$ with asymptotically normal distribution and, under the model, $J(\hat{\beta}^{GMM}) = O_p(1)$ with χ^2_{m-p} asymptotic calibration, where $J(\beta) \equiv J(\beta; \beta)$ (Hansen, 1982; Newey and McFadden, 1994; Gallant and White, 1988).

To describe a GEL procedure let $\rho(\upsilon)$ be a function of a scalar υ that is convex on an open interval $\mathcal{V}$ containing zero. A GEL estimator for the parameter β defined by the moment condition(1) is given by

$$\hat{\beta}^{GEL} = \arg\max_{\beta \in B} \sum_{i=1}^{n} \rho(\tau(\beta)' g_i(\beta))/n, \ \tau(\beta) = \arg\min_{t \in T(\beta)} \sum_{i=1}^{n} \rho(t' g_i(\beta)), \ T(\beta) = \{t : t' g_i(\beta) \in \mathcal{V}\}.$$

The empirical likelihood estimator is obtained when $\rho(\upsilon) = -\log(1-\upsilon)$, exponential tilting when $\rho(\upsilon) = \exp(\upsilon) - 1$, and the continuous updating estimator when $\rho(\upsilon) = \upsilon + \upsilon^2/2$. The GEL estimator has some interesting frequentist properties (see Newey and Smith, 2004 and Ragusa, 2006b). Central to the understanding of the results of this paper is the fact that the GEL procedure delivers efficient estimators of the c.d.f. of x under the moment condition. This class of estimators is defined as $\widehat{F(x)} = \sum 1(x_i \leq x)\varphi_i$ where $\varphi_i = \rho_1(\tau' g_i(\beta))/\sum_{i=1}^{n} \rho_1(\tau' g_i(\beta))$, $\rho_1(\upsilon) = \partial\rho(\upsilon)/\partial\upsilon$ can be interpreted as the estimated probability of the observations x_i under the model.

Fully Bayesian approaches to inference in semiparametric models boil down to finding an approximate likelihood consistent with the model under consideration and with the Bayesian learning mechanism. Mixed approaches try to justify the use of common criterion functions, such as $J(\cdot, \cdot)$, as a central strategy for obtaining a Bayesian likelihood.

Chamberlain and Imbens (2003) extend to semiparametric models the Bayesian bootstrap (Rubin, 1981).[1] In moment condition models, Bayesian bootstrap consists of solving a weighted version of the sample moment equation where the weights are sets of i.i.d. ex-

[1]Bayesian bootstrap in nonparametric settings has been considered by Ferguson (1973, 1974) and Gasparini (1995). Hahn (1997) studies the frequentist properties of the BB for the quantile regression case and Lancaster (1994) applies BB to the analysis of choice based sampling.

ponential random variables $V_i^{(s)}(i = 1, \ldots, n)$:

$$\sum_{i=1}^{n} V_i^{(s)} g(x_i, \beta^{(s)}) = 0, \quad (l = 1, \ldots, S).$$

The procedure gives S independent draws from a posterior distribution obtained by assuming (an improper) Dirichlet prior on the space of distributions. There are two main problems with this approach. First, it is not clear how to incorporate prior knowledge about β. The prior on β is implicitly elicited by the choice of the Dirichlet prior. Second, drawings are obtained by solving a potentially high-dimensional nonlinear set of equations. When $m > p$, one needs to augment the parameter vector and the moment functions. This augmentation takes place outside the model and hence the resulting inference is arbitrary.

For an alternative likelihood function in Bayes' theorem, Kim (2002) proposes using a transformation of the efficient GMM objective function, namely $\exp\{-J(\beta)/2\}$. This procedure is only justified asymptotically and is not based on formal Bayesian calculus.

Lazar (2003) proposes replacing the likelihood in the formula for the posterior with the empirical likelihood

$$L(\mathrm{x}|\beta) = \left\{ \max_{\pi} \prod_{i=1}^{n} \pi_i, \text{ subject to } \sum_{i=1}^{n} \pi_i g(x_i, \beta) = 0; \sum_{i=1}^{n} \pi_i = 1 \right\}.$$

A theoretical obstacle for using EL is that the nuisance parameters π_i $(i = 1, \ldots, n)$ are maximized over, giving an estimated likelihood. Even if estimated likelihoods have been used to approximate marginal likelihoods, there is no reason to expect that a profiled likelihood will behave like a marginal likelihood. A proper and fully Bayesian approach would require eliciting a prior distribution for π conditional on β and integrating out π with respect to this prior to obtain a marginal posterior.

Schennach (2005) recognizes the theoretical limitations of EL-based likelihoods and successfully seeks remedies by adopting a formal Bayesian procedure in which the nuisance parameters are assumed to grow to infinity. The procedure obtains the Bayesian Exponentially Tilted Empirical Likelihood (BETEL), defined as:

$$L(\mathrm{x}|\beta) = \prod_{i=1}^{n} \frac{\exp(\tau' g_i(\beta))}{\sum_{i=1}^{n} \exp(\tau' g_i(\beta))}, \quad \tau = \arg\min_{t} \sum_{i=1}^{n} \exp(t' g_i(\beta)).$$

As opposed to Schennach (2005) in this paper we do not require the nuisance parameters to grow to infinity. Rather nonparametric assumptions are made on the support of the distribution. These assumptions greatly simplify the problem and deliver a transparent and intuitive derivation of a class of likelihoods that are consistent with Bayesian calculus. As shown in the next section, these likelihoods are connected with EL, ET, and, in general, with GEL methods.

3 Bayesian Likelihood

The analysis assumes that the distribution of the random vector x belongs to $\mathcal{F}_\theta = \{F_\theta : \theta \in \Theta\}$, the class of discrete distributions. The distributions F_θ have finite support, $\Pr(x = a_j) = \theta_j$, $(j = 1, \ldots, J)$, where θ_j is the jth element of $\theta \in \Theta$ and Θ is the unit simplex in $\mathbb{R}^J$. In the Bayesian bootstrap, the discreteness of x is a property of the posterior distribution.[2] In the present setting, the restriction to the class of discrete distributions is made for convenience. Since J can be large and all data are observed discretely, assuming discreetness is no real restriction (Rubin, 1981 page 133).

When the model is restricted to $\mathcal{F}_\theta$, the moment condition can be rewritten as

$$\sum_{j=1}^{J} g(a_j, \beta_0)\theta_j = 0.$$

Let $n_j = \sum_{i=1}^{n} 1(x_i = a_j)$ be the number of sample observations equal to a_j. The likelihood of $\mathrm{x} = (x_1, \ldots, x_n)$ can then be written as

$$L(\mathrm{x}|\beta) = \prod_{j=1}^{J} \theta_j^{n_j}.$$

The weights θ_j $(j = 1, \ldots, J)$ can be apportioned into observation-specific weights $\omega_i > 0$ $(i = 1, \ldots, n)$ by requiring that the sum of the ω_i for all the observations equal to a_j is θ_j. Formally, the ω_i $(i = 1, \ldots, n)$ are implicitly defined by J equations: $\theta_j = \sum_{i=1}^{n} 1(x_i = a_j)\omega_i$. These equations do not pin down a unique value for the ω_i, as there are many ways to assign values to ω_i and still satisfy the J equations. Let $\theta_j(i)$ be the probability attached by F_θ to

[2] A draw from a Dirichlet process is a distribution that places its probability mass on a countably infinite subset of the underlying sample space.

the support point a_j that corresponds to observation i, $\theta_j(i) = 1(x_i = a_j)\Pr(x = a_j)$, and let $n_j(i) = \sum_{j=1}^{J} 1(x_i = a_j)$. The weights can be uniquely identified by setting $\omega_i = \theta_j(i)/n_j(i)$ from which it follows that

$$\prod_{i=1}^{n} \omega_i = \prod_{i=1}^{n} \theta_j(i)/n_j(i) = \prod_{i=1}^{n} \theta_j(i) / \prod_{i=1}^{n} n_j(i) \propto \prod_{j=1}^{J} \theta_j^{n_j}.$$

The population moment condition can be rewritten in terms of the observation-specific weights:

$$\sum_{j=1}^{J} \theta_j g(a_j, \beta_0) = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{J} 1(x_i = a_j) g(a_j, \beta_0) + \sum_{j=1}^{J} 1(n_j = 0)\theta_j g(a_j, \beta_0).$$

If the support of x has been explored by the data, then $\sum_{j=1}^{J} 1(n_j = 0)\theta_j g(a_j, \beta_0) = 0$, and the moment condition becomes

$$0 = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{J} 1(x_i = a_j) g(a_j, \beta_0) = \sum_{i=1}^{n} \omega_i g(x_i, \beta_0).$$

The assumption that all possible distinct values of x have been observed is questionable. This assumption is also made by the BB. Imbens and Chamberlain (2003) justify it from both a theoretical and computational point of view. Rubin (1981), however, discusses its potential pitfalls. If $\Pr(x \geq x_{(n)}) \neq 0$, where $x_{(n)}$ denotes the nth order statistics, then assuming that $\Pr(x \geq x_{(n)}) = 0$ will have an impact on the resulting inference. The impact of this assumption is addressed in Section 5 where the coverage properties of the resulting posteriors are examined.

When the distribution of x belongs to $\mathcal{F}_\theta$ and all the possible values of x have been observed, the model can be expressed as a likelihood function, $\prod_{i=1}^{n} \omega_i$, and a moment condition, $\sum_i \omega_i g(x_i, \beta_0) = 0$, both expressed in terms of observation-specific weights, ω_i $(i = 1, \ldots, n)$.

Since inference is about β, the observation specific weights are nuisance parameters. In a formal Bayesian framework, nuisance parameters are eliminated through integration. If the joint prior on (ω, β), $\pi(\omega, \beta)$, is absolutely continuous with respect to the Lebesgue measure,

then the marginal posterior of β is

$$\pi(\beta|\mathrm{x}) \propto \left[\int L(\mathrm{x}|\beta,\omega)\pi(\omega|\beta)d\omega\right]\pi(\beta)$$

where $\pi(\omega,\beta)$ has been partitioned as $\pi(\omega,\beta)=\pi(\omega|\beta)\pi(\beta)$. While assuming an absolutely continuous prior distribution for β is not restrictive, it is important to consider the more general case where the distribution of ω given β is not necessarily absolutely continuous. A representation for the marginal posterior of β that takes into account this possibility is

$$\pi(\beta|\mathrm{x}) \propto \left[\int L(\mathrm{x}|\beta,\omega)dP_{\omega\beta}\right]\pi(\beta),$$

where $P_{\omega,\beta}$ denotes the joint prior measure on (ω,β), and $dP_{\omega,\beta}$ has been partitioned as $dP_{\omega,\beta} = dP_{\omega\beta}\times dP_{\beta}$. Given that in our setting the likelihood $\prod_{i=1}^{n}\omega_i$ does not depend directly on β, the marginal posterior becomes

$$\pi(\beta|\mathrm{x}) \propto \left\{\int\left[\prod_{i=1}^{n}\omega_i\right]dP_{\omega\beta}\right\}\pi(\beta). \tag{2}$$

For the purpose of inference about β, two distributions must be elicited: (i) an unconditional prior distribution forβ, and (ii) a distribution for ω given β. Elicitation of $\pi(\beta)$ is in the realm of Bayesian analysis and the usual considerations on prior selection apply (Berger, 1985). Given β, eliciting a prior distribution for ω amounts to choosing the way in which the statistical information in the moment condition is conveyed into the posterior distribution. Our innovation concerns the specification of this conditional prior distribution: using the Kullback-Leibler information-theoretic measure we derive a distribution $P_{\omega\beta}$ that fully accounts for the information in the model but remains analytically tractable.

3.1 Minimum Kullback-Leibler conditional distribution

This section shows how the information contained in the moment condition can be translated into a conditional measure $P_{\omega\beta}$ that is then used to integrate out the nuisance parameters of the likelihood, as specified in (2). The starting point is to specify a prior measure P_{ω} on ω that does not account for the information in the model. Let $b_x(\omega,\theta)=\sum_{i=1}^{n}\omega_i g(x_i,\beta)$.

Consider the following set of measures

$$\mathcal{P}(\beta) = \left\{ \mu \left| \int b_x(\omega, \theta) d\mu = 0, \ \int d\mu = 1, \ \mu \ll P_\omega \right. \right\}$$

The set $\mathcal{P}(\beta)$ contains the (P_ω-absolutely continuous) measures that satisfy an average form of the moment conditions. Among the elements of $\mathcal{P}(\beta)$, we choose the measure that minimizes the Kullback-Leibler information number with respect to P_ω, that is the solution to the following optimization problem (Csiszar, 1975)

$$P_{\omega\beta} = \arg \min_{\mu \in \mathcal{P}(\beta)} D(\mu, P_\omega), \quad D(Q, P) = \begin{cases} \int \log \left(\frac{dQ}{dP} \right) dQ & \text{if } Q \ll P \\ +\infty & \text{otherwise} \end{cases} \tag{3}$$

The Kullback-Leibler information number for two probability measures Q and P, $D(Q, P)$, is always non negative and $D(Q, P) = 0$ iff $P = Q$ a.s. If such $P_{\omega\beta}$ exists, the convexity of $\mathcal{P}(\beta)$ guarantees its uniqueness since $D(Q, P)$ is a strictly convex functional in Q.[3] The probability measure $P_{\omega\beta}$ is referred to as the I-projection of P_ω into $\mathcal{P}(\beta)$.

Minimization problems of the type in (3) play a basic role in information-theoretic approaches to statistics. In the Bayesian literature, if P_ω were the natural invariant non-informative prior for the problem, $\max_{Q \in \mathcal{P}(\beta)} -D(Q, P_\omega)$ would be the equivalent to the maximum entropy distribution in the presence of partial prior information (Jaynes, 1968).

In our setting the intuition behind $P_{\omega\beta}$ is that it is the distribution that contains the information about $b_x(\omega, \beta) = 0$ but has the largest entropy with respect to a given (unconditional) prior distribution on the observation-specific weights.

The following assumptions will be maintained throughout the paper. Let $S_\rho(\beta_0)$ be an open neighborhood of β of radius$\rho > 0$.

Assumption 1. *The measure P_ω has a product representation, $P_\omega = P_{\omega,1} \times \ldots \times P_{\omega,n}$*

Assumption 2. *For $\beta \in S_\rho(\beta_0)$*

[3] As in Csiszar (1975), in addition to the arithmetic of the extended reals, the following conventions regarding infinity are adopted in keeping with measure-theoretically consistent operations:

$$\log 0 = -\infty, \quad \log \frac{a}{0} = +\infty, \quad 0(\pm\infty) = 0$$

(a) The set

$$T_{P_\omega}(\beta) = \left\{ \tau : \int \exp(\tau b_x(\omega, \beta)) dP_\omega < \infty \right\}$$

is open in $\mathbb{R}^m$.

(b) the $n \times m$ *matrix* $G(\beta) = (g_1(\beta)\, g_2(\beta)\, \cdots g_n(\beta))'$ *has full column rank.*

Assumption 1 states that under the prior distribution P_ω the observation probability weights are independent, each with distribution $P_{\omega,i}$. The BB makes the same assumption on the prior probabilities. Assumption 2 is technical and is required to show the existence of the I-projection of P_ω on $\mathcal{P}(\beta)$.

The following proposition gives conditions for the existence and the form of the I-projection. The proof is based on Theorem 3.4 in Csiszar (1975). Let

$$\mathcal{P}(\beta; a) = \left\{ \mu : \int b_x(\omega, \beta) d\mu = a,\, a \in \mathbb{R}^m \right\},$$

A_m be the set of points $a \in \mathbb{R}^m$ for which $\mathcal{P}(\beta, a)$ contains some μ such that $D(\mu, P_\omega) < \infty$, and $\bar{A}_{P_\omega}$ the interior of A_{P_ω}.

Proposition 1. *If Assumptions 1-2 are satisfied, then* A_{P_ω} *is non-empty, and if* $0 \in \bar{A}_{P_\omega}$ *then the I-projection of* P_ω *on* $\mathcal{P}(\beta)$ *exists and it has the form*

$$P_{\omega\beta} = \begin{cases} c_x(\beta) \cdot \exp(t' b_x(\omega, \beta)) P_\omega & \text{if } \omega \notin N \\ 0 & \text{if } \omega \in N \end{cases} \tag{4}$$

where

$$t = \arg \min_{\tau \in T_{P_\omega}} \int \exp(\tau' b_x(\omega, \beta)) dP_\omega, \quad c_x(\beta) = \left[\int \exp(t' b_x(\omega, \beta)) dP_\omega \right]^{-1}$$

and N *has* $\mu(N) = 0$ *for every* $\mu \in \mathcal{P}(\beta)$.

In the case considered here, the existence of the I-projection cannot be established by invoking closeness of the set $\mathcal{P}(\beta)$ with respect to the variational distance and using Theorem 2.1 in Csiszar (1975). The function $b_x(\omega, \beta)$ is unbounded, and closure under the variational distance is, in this case, inappropriate.

The conditional measure (4) can be rewritten as

$$P_{\omega\beta} = \exp(t' b_x(\omega, \beta) - \log c_x(\beta)) P_\omega$$

Let $u_i(\beta, t) = t' g_i(\beta)$ and notice that $t' b_x(\omega, \beta) = \sum_{i=1}^n \omega_i u_i(\beta, t)$. Thanks to the special structure of $b_x(\omega, \theta)$, the I-projection $P_{\omega\beta}$ is a product measure with components that are proportional to the measures $P_{\omega,i}$. This in turn implies that under the measure $P_{\omega\beta}$ the observation-specific weights are independent.

Proposition 2. *Under Assumption 1-2, the ω_i $(i = 1, \ldots, n)$ are independent under the I-projection measure $P_{\omega\beta}$, each ω_i having distribution $P_{\omega\beta,i}$ given by*

$$P_{\omega\beta,i} = \exp(\omega_i u_i(\beta, \tau) - k_x(u_i(\beta, \tau)) P_{\omega,i}$$

where $k_x(\upsilon_i) = \log \int \exp(\omega \upsilon_i) dP_{\omega,i}$ and $\tau = \arg\min_{t \in T_{P_\omega}} \sum_{i=1}^n \int \exp(\omega_i u_i(\beta, t)) dP_{\omega,i}$.

The result of Proposition 2 constitutes the building block for the derivation of the semi-parametric likelihoods. The measures $P_{\omega\beta,i}$ will be treated as priors on ω_i with respect to which the inner integration in (2) is evaluated. Notice that the $P_{\omega\beta,i}$ $(i = 1, \ldots, n)$ are consistent with the way the probabilities θ_j $(j = 1, \ldots, J)$ are apportioned into weights ω_i $(i = 1, \ldots, n)$, since $P_{\omega\beta,i}(B) = P_{\omega\beta,k}(B)$ for any set B whenever $x_i = x_k$, $(i = k)$. The functional form of the distributions $P_{\omega\beta,i} (i = 1, \ldots, n)$ relates the initial prior distribution $P_{\omega,i}$ to $P_{\omega\beta,i}$ through the term $k_x(\cdot)$, the cumulant generating function of the random variable ω_i.

3.2 Integration

The next step in deriving a Bayesian likelihood consists in carrying out the integration with respect to $P_{\omega\beta} = P_{\omega\beta,1} \times \cdots \times P_{\omega\beta,n}$:

$$\pi(\beta|\mathrm{x}) \propto \left\{ \int \left[\prod_{i=1}^n \omega_i \right] d\left(P_{\omega\beta,1} \times \cdots \times P_{\omega\beta,n} \right) \right\} \pi(\beta)$$

It turns out that the integration step in the posterior above is tractable, as the following proposition makes clear.

Proposition 3. *Integrating the (unconstrained) likelihood with respect to $P_{\omega\beta}$ gives a posterior*

$$\pi(\beta|\mathrm{x}) \propto \bar{L}(\mathrm{x}|\beta)\pi(\beta),\ \bar{L}(\beta|\mathrm{x}) = \prod_{i=1}^{n} k_{x,\partial}(u_i(\beta,\tau))$$

where

$$k_{x,\partial}(\upsilon_i) = \int \omega \exp(\omega\upsilon_i - k_x(\upsilon_i))dP_{\omega,i} / \sum_{i=1}^{n} \int \omega \exp(\omega\upsilon_i - k_x(\upsilon_i))dP_{\omega,i}$$

$\upsilon_i = u_i(\beta,\tau)$ *and* $\tau = \arg\min_{t\in T_{P_\omega}} \sum_{i=1}^{n} \int \exp(\omega u_i(\beta,t))dP_{\omega,i}$.

This last result shows that the integrate likelihood function, $\bar{L}(\beta|\mathrm{x})$, is the product of n terms of the form $k_{x,\partial}(\upsilon_i)$. Notice that the numerator of $k_{x,\partial}(\upsilon_i)$ is the first derivative of the cumulant generating function of ω under the prior $P_{\omega,i}$:

$$\begin{aligned}
&\frac{d}{d\upsilon_i} \log \int \exp(\omega\upsilon_i)dP_{\omega,i} \\
&= \left(\int \exp(\omega\upsilon_i)dP_{\omega,i}\right)^{-1} \int \omega \exp(\omega\upsilon_i)dP_{\omega,i} \\
&= \int \omega \exp(\omega\upsilon_i - k_x(\upsilon_i))dP_{\omega,i}.
\end{aligned}$$

The parameter τ also depends on $k_x(\upsilon_i)$. It turns out that for many initial prior distributions P_ω, both $k_x(\upsilon_i)$ and $k_{x,\partial}(\upsilon_i)$ have a simple mathematical form that allows expressing τ and the posterior distribution as functions of $u(\beta,\tau)$ that do not involve integrals.

4 Connections with GEL methods

What is the form of $\bar{L}(\beta|\mathrm{x})$ when the initial distribution P_ω belong to a given class? In which cases the functional form is tractable from a computational point of view? Is there a relationship between the likelihoods obtained in Section 3 and the ones given in the literature and briefly discussed in Section 2? This section seeks to answer these questions.

The dependence of the integrated likelihoods on the initial prior is through the cumulant generating function of the observation specific weights (under P_ω). Judiciously choosing the initial prior in such a way that $k_x(\upsilon_i)$ and $k_{x,\partial}(\upsilon_i)$ have a simple mathematical form allows us to connect the likelihoods to GEL. As shown by the following proposition, eliciting

a Normal, Poisson, and Gamma initial prior for ω gives likelihoods of very simple and recognizable form.

Proposition 4. *If under the initial prior distribution the* ω_i $(i = 1, \ldots, n)$ *are:*
a) iid with $\omega_i \sim N(1, \sigma^2)$ *, then* $\bar{L}(\mathrm{x}|\beta)$ *has the following form*

$$\bar{L}(\mathrm{x}|\beta) = \prod_{i=1}^{n} 1 + \sigma^2 u_i(\beta, \tau)/2 / \sum_{i=1}^{n} \left(1 + \sigma^2 u_i(\beta, \tau)/2\right), \tag{5}$$

where $\tau = \arg\min_t \sum_{i=1}^{n} (1 + \sigma^2 u_i(\beta, t)/2)^2$.
b) iid with $\omega_i \sim Poisson(1)$*, then* $\bar{L}(\mathrm{x}|\beta)$ *has the following form*

$$\bar{L}(\mathrm{x}|\beta) = \prod_{i=1}^{n} \exp(u_i(\beta, \tau)) / \sum_{i=1}^{n} \exp(u_i(\beta, \tau)), \tag{6}$$

where $\tau = \arg\min_t \sum_{i=1}^{n} \exp(u_i(\beta, t))$.
c) iid with $\omega_i \sim Gamma(1, \varsigma)$,$\varsigma > 0$*, then* $\bar{L}(\mathrm{x}|\beta)$ *has the following form*

$$\bar{L}(\mathrm{x}|\beta) = \prod_{i=1}^{n} (1 - \varsigma u_i(\beta, \tau))^{-1} / \sum_{i=1}^{n} (1 - \varsigma u_i(\beta, \tau))^{-1}, \tag{7}$$

where $\tau = \arg\min_{t \in T_\gamma(\beta)} - \sum_{i=1}^{n} \log(1 - \varsigma u_i(\beta, t))$ *and* $T_\varsigma(\beta) = \{t : \max_{i<n} u_i(\beta, t) < 1/\varsigma\}$.

In each case of Proposition 4, and of Proposition 5 below, the function to be minimized in order to obtain τ is strictly convex, making the likelihood computationally tractable. However, for likelihood in (7), and the other likelihoods of Proposition 5, the optimization must be carried on a restricted set.

Proposition 4 shows that there is a close relationship between $\bar{L}(\mathrm{x}|\beta)$ obtained in Section 3 and GEL. For a given $\beta \in \mathcal{B}$, the likelihood is the product of the GEL weights for a specific criterion function $\rho(\cdot)$. A $N(1, 1)$ initial prior corresponds to a likelihood that is the product of the weights of the CUE, while a $Gamma(1, 1)$ prior yields a likelihood that is the product of EL weights. Similarly, when the prior is $Poisson(1)$ then $\bar{L}(\mathrm{x}|\beta)$ is the product of the weights of ET. This last case corresponds to the BETEL in Schennach (2005).

Poisson, Normal and Gamma are not the only distributions for which the likelihood displays a relationship with the GEL. The next proposition shows that there exists an initial prior distribution such that the resulting $\bar{L}(\mathrm{x}|\beta)$ is the product of GEL weights when

the criterion function corresponds to the Cressie-Read (CR) divergence, that is $\rho(\upsilon) = (1+\gamma\upsilon)^{(\gamma+1)/\gamma}/(\gamma+1)$.[4]

Proposition 5. *There exists an initial prior distribution on* ω_i $(i = 1, \ldots, n)$ *such that*

$$\bar{L}(\mathrm{x}|\beta) \propto \prod_{i=1}^{n} \frac{(1+\gamma u_i(\beta,\tau))^{1/\gamma}}{\sum_{i=1}^{n}(1+\gamma u_i(\beta,\tau))^{1/\gamma}}, \tag{8}$$

where $\tau = \arg\min_{t\in T_\gamma(\beta)} -\sum_{i=1}^{n}(1+\gamma u_i(\beta,t))^{(\gamma+1)/\gamma}$ *and* $T_\gamma(\beta) = \{t : \max_i u_i(\beta,t) < -1/\gamma\}$,$\gamma < -1$.

The correspondence between $\bar{L}(\mathrm{x}|\beta)$ and the CR criterion function established in Proposition 5 holds only for $\gamma < -1$. However, Proposition 5 still includes important cases such as the Hellinger distance ($\gamma = -3/2$).

Notice that the likelihoods obtained by assuming Normal and Gamma priors as in Proposition 4(a) and 4(c) are more general than EL- and CUE-based likelihoods since they allow for parameters, σ and ς, that control the spread of the prior distributions. When $\sigma \to 0$ and $\varsigma \to 0$ the information in the moment condition is annihilated and the likelihoods become concentrated at n^{-n} regardless of the value of β. The likelihood in (6) does not admit a variance parameter, because in the Poisson case variance and mean are parameterized by the same parameter and hence it is impossible to control the variance without affecting the mean of the distribution.

5 Validity

Is inference based on $\pi(\beta|\mathrm{x}) \propto \bar{L}(\mathrm{x}|\beta)\pi(\beta)$ valid? Monahan and Boos (1992) deem a likelihood valid, in the sense of giving valid posterior inference, if: (i) is based on the conditional density of the data given the parameter of interest; (ii) is supported by probability calculus. Strictly adhering to *i)* would preclude considering Bayesian analysis in nonparametric and semiparametric settings where, by definition, the conditional distribution of the data given

[4] As shown in Newey and Smith (2004), minimization of $\sum_{i=1}^{n} h(p_i)/n$ subject to $\sum_{i=1}^{n} p_i g_i(\beta) = 0$, $\sum_{i=1}^{n} p_i = 0$ over multinomial distributions putting masses $(p_1, \ldots, p_n)$ on $(x_1, \ldots, x_n)$ when $h(x) = [(nx)^{\gamma+1} - 1]/\gamma(\gamma+1)$ is related to solving $\min_t \sum_{i=1}^{n} \rho(t' g_i(\beta))/n$ where $\rho(\upsilon) = (1+\gamma\upsilon)^{(\gamma+1)/\gamma}/(\gamma+1)$. This form of $h(x)$ was first studied by Cressie and Read (1984) and it is often referred to as the Cressie-Read divergence.

the parameter is unavailable. The second requirement is to be taken seriously. By probability calculus we mean the application of Bayes theorem and the integration of nuisance parameters with respect to a prior distribution.

In what follows we investigate to what extent $\bar{\pi}(\beta|\mathrm{x})$ provides valid posterior inference. We will rely on the validity by coverage concept of Monahan and Boos (1992). If a posterior is valid by coverage then, for every continuous prior distribution $\pi(\beta)$, the quantity $a(\mathrm{x},\beta) = \int_{-\infty}^{\beta} \bar{\pi}(u|\mathrm{x})du$ is unconditionally distributed as $U(0,1)$. To see it, let $Q(u) = \int_{-\infty}^{u} f(\mathrm{x}|\beta)\pi(\beta)d\beta$, $Q(y)^{-1} = \inf\{u : Q(u) \geq y\}$, and $a(\mathrm{x},y)^{-1} = \inf\{u : a(x,u) \geq y\}$. Then, if $\pi(\beta|\mathrm{x}) \propto f(\mathrm{x}|\beta)\pi(\beta)$, we have

$$\begin{aligned} & \int\int 1(a(\mathrm{x},\beta) \leq z) f(\mathrm{x}|\beta)\pi(\beta)d\beta dx \\ = & \int \left[\int 1(a(\mathrm{x},\beta) \leq z)\pi(\beta|x)d\beta\right] f(\mathrm{x})dx \\ = & \int Q(a(\mathrm{x},z)^{-1}) f(\mathrm{x})d\mathrm{x} \\ = & \ z \end{aligned}$$

where the last equality follows from the fact that $Q(a(z)^{-1}) = z$ for any x.

When the posterior is based on the (scaled) true conditional density, $f(\mathrm{x}|\beta)$, $a(\beta) \sim U(0,1)$, the deviation from $U(0,1)$ indicates that the posterior probability regions have the wrong coverage probabilities. Operationally, the procedures works as follows. First, generate $\beta^{(j)}$, $j = (1,\ldots,R)$ independently from $\pi(\beta)$. Then, conditionally on each $\beta^{(j)}(j = 1,\ldots,R)$ generate the data according to $f(\cdot|\beta^{(j)})$, obtaining $\mathrm{x}^{(j)}(j = 1,\ldots,n)$. Finally, compute $a(\beta^{(j)}) = \int_{-\infty}^{\beta^{(j)}} \bar{\pi}(\mathrm{x}^{(j)}|\beta)\mathrm{d}\beta$.

Assessing validity by coverage boils down to calculating test statistics to quantify their deviation from uniformity, and hence detecting bad coverage.

We consider two simple models. The first model is a quantile restriction. This case is particularly interesting because it shows that in certain instances, i.e., in a model defined by bounded moment functions, the choice of the prior distribution for the nuisance parameters is immaterial. The second is an overidentified normal location model. This model is interesting because often the moment condition models considered in economics are overidentified.

5.1 Quantile

The qth quantile of the real valued random variable X with distribution function $F(x) = \Pr(X \leq x)$ is defined as $\beta_q = \inf\{x : F(x) \geq q\}$. The qth quantile satisfies the following moment condition

$$E\left[\delta(x, \beta_q) - q\right] = 0 \tag{9}$$

where $\delta(x, \beta) = 1(x \leq \beta)$. For the population median, $q = 1/2$, analysis of the validity by coverage of (7) has been carried out by Lazar (2003), who also consider the nonparametric likelihood based on the binomial character of the empirical distribution function proposed by Jeffreys (1967, p.211-2)

$$\begin{pmatrix} n \\ k(\beta_{1/2}) \end{pmatrix} (1/2)^n \tag{10}$$

where $k(\beta) = \sum_{i=1}^{n} \delta(x_i, \beta)$.

We note that, in this case, the integrated likelihoods derived in Section 4 have a closed-form solutions. Consider the likelihood resulting from multiplying the weights of EL and the likelihood resulting from multiplying the weights of the ET. These two likelihoods are equivalent up to a proportionality constant, as the next result shows.

Appendix B derives expressions for the likelihoods given in Proposition 4 and Proposition 5. In particular, it is shown that 7, 6, 8 and 5 are equivalent up to a proportionality constant that is

$$\begin{aligned} \bar{L}^{EL}(\mathrm{x}|\beta_q) \propto \bar{L}^{ET}(\mathrm{x}|\beta_q) \propto \bar{L}^{CR}_{\gamma}(\mathrm{x}|\beta_q) \propto \bar{L}^{CUE}(\mathrm{x}|\beta) \\ \propto \left(\frac{q}{k(\beta_q)}\right)^{k(\beta_q)} \left(\frac{(1-q)}{1-k(\beta_q)}\right)^{n-k(\beta_q)} \end{aligned} \tag{11}$$

Equation 11 implies that Bayesian inference based either on $\bar{L}^{EL}(\mathrm{x}|\beta_q)$, $\bar{L}^{ET}(\mathrm{x}|\beta_q)$, $\bar{L}^{CR}_{\gamma}(\mathrm{x}|\beta_q)$ and $\bar{L}^{CUE}(\mathrm{x}|\beta)$ will be equivalent.[5] Notice that, as for the nonparametric likelihood in 10, the support of 11 coincides with the set of observed values $\mathrm{x} = (x_1, \ldots, x_n)$ and that the like-

[5] Ragusa (2006a) shows that MLE estimators based on likelihood like the ones in 11 and relative to parameters defined through smooth moment restrictions are third order efficient after bias correction. The fact that when the model is $E[\delta(x, \beta_q) - q] = 0$ the likelihood based on EL, ET, and on CR are proportional seems to suggest that some form of higher order equivalence may apply to nonsmooth setting as well.

lihood is constant on $\beta_q \in [x_{i-1}, x_i)$, $(i = 1, \ldots, n)$. These features of 11 and 10 have computational implications that lead to an easy method for calculating quantiles of the posteriors. Let $\bar{L}_k$ denote the value of the likelihood on each interval $(x_{k+1} - x_k)$, $(k = 1, \ldots, n-1)$. Then the percentiles of the posterior distribution are given by

$$P_{\pi(\beta|\mathrm{x})}(\beta \leq p) = \frac{\nabla L_1 + \nabla L_2}{\nabla L}$$

where

$$\nabla L_1 = \sum_{k=1}^{k_p} (x_{k+1} - x_k)\bar{L}_k \left(\int_{-\infty}^{x_{k+1}} \pi(\beta)d\beta - \int_{-\infty}^{x_k} \pi(\beta)d\beta \right)$$

$$\nabla L_2 = (p - x_{k_p})\bar{L}_k \left(\int_{-\infty}^{x_{k_p}} \pi(\beta)d\beta - \int_{-\infty}^{p} \pi(\beta)d\beta \right)$$

$$\nabla L = \sum_{k=1}^{n-1} (x_{k+1} - x_k)\bar{L}_k \left(\int_{-\infty}^{x_{k+1}} \pi(\beta)d\beta - \int_{-\infty}^{x_k} \pi(\beta)d\beta \right)$$

and $k_p = \sup\{k : x_k < p\}$.

Figure 1 plots the posterior distribution for the median of a sample of ten observations based on (10) and (11).[6] The posterior is plotted for three proper priors: $N(0, \sqrt{2})$, $N(0, 1)$ and $U(-3, 3)$, as well as for the the improper diffuse prior $\pi(\beta) \propto 1$. In each graph, the continuous line plots the posterior distribution based on (11) and the dashed line plots the posterior based on (10). The curves describing the (proper) prior distributions are also plotted. The band in the top portion of each graph represents the 95% confidence interval based on asymptotic calculations. The posterior distribution based on the Jeffreys nonparametric likelihood is virtually indistinguishable from the one based on (11). There are minimal differences on the mass of probability the two methods put on the tail, with the Jeffreys' likelihood-based posterior putting more mass on the tails.

Next we calculate $a(\beta)$ when $(x_1, \ldots, x_n)$ are iid $N(0, 1)$, $n = 20$ and $\beta \sim N(0, \sigma_\pi^2)$, $\sigma_\pi^2 = \{.1, 3, 12\}$. Figure 2 plots the quantiles of $a(\beta)$ against the quantiles of the distribution of $U(0, 1)$ while Figure 3 plots the histogram of $a(\beta)$ for the three prior distributions. The

[6] The ten observations are

x = {0.507875807673672, 4.43238097646257, −1.69965269919334, −0.753758224087492, 0.580532998124808, 1.45526401607326, 1.64006931129037, 3.25264997161796, 0.572136677346761, 7.15694257504198}

plots are based on 10,000 replications. When the prior is very tight, both the EL, ET, CUE and CR class of likelihoods and the Jeffreys' likelihood are substantially valid by coverage. When the prior is more disperse, the nonparametric likelihoods are invalid by coverage, as clearly showed by the quantiles-to-quantiles plot and the histograms.

5.2 Overidentified Location Scale Model

Here we consider an overidentified model. The parameter of interest is defined by

$$E[g(x,\beta_0)]=0, \quad g(x,\beta)=\begin{bmatrix} x-\beta \\ (x-\beta)^2-1 \end{bmatrix}.$$

and $x \sim N(0,1)$. In this case, the value of the parameter that solves the moment condition is $\beta_0 = 0$. Bayesian inference about β is based on the posterior

$$\pi^{Bayes}(\beta|\mathrm{x}) = \frac{\prod_{i=1}^{n} f(x_i|\beta)\pi(\beta)}{\int \prod_{i=1}^{n} f(x_i|\beta)\pi(\beta)d\beta} \tag{12}$$

where $f(x_i|\beta) \propto e^{-(\beta-x)^2/2}$. For simplicity, we calculate $a(\beta)$ for the semiparametric likelihoods based on EL, ET, and CUE. Three sets of priors are considered: (i) $\beta \sim U(-1,1)$; (ii) $\beta \sim U(-2,2)$; *iii)* $\beta \sim U(-5,5)$. The choice of uniform priors is for convenience, since in this case the numerical integrations required to rescale the likelihoods can be carried out on compact intervals. The nonparametric likelihoods are calculated by solving their respective optimization problems that define the parameter τ.

Figures 5 to 7 show quantiles-to-quantiles plots of the criterion $a(\beta)$ for the set of priors for sample sizes of $n = \{20, 50, 100\}$. It is immediately apparent that the likelihood based on the CUE is invalid by coverage for any sample size considered here. Also, the quantiles of $a(\beta)$ and the quantiles of the uniform distribution do not get closer as the sample size gets larger. The quantiles-to-quantiles plots for the EL and ET based likelihoods show instead that both these likelihoods are approximately valid by coverage. To see whether the distribution of $a(\beta)$ is $U(0,1)$ a Kolmogorov Smirnov (KS) test can be employed in each case. Figure 4 plots the p-values of the KS-statistics for the posterior based on EL, ET and CUE. The first horizontal line from the bottom denotes 5% significance; the second horizontal line denotes 10% significance. The plots clearly show that for all the priors the

distribution of $a(\beta)$ is statistically indistinguishable from that of $U(-1,1)$ when $n = 50$.

6 Conclusion and future work

This paper proposes a new class of semiparametric likelihoods for Bayesian inference. These likelihoods are useful when the only information about the parameter of interest is expressed in terms of a moment condition. This paper contributes a method for incorporating the information about the parameter into a likelihood function. This is achieved through the integration of nuisance parameters with respect to a tilted distribution that is consistent with the moment condition. The result is a semiparametric likelihood whose features depend on the initial prior elicited on the nuisance parameters.

We are able to link these likelihoods to GEL methods. In particular, we obtain likelihoods related to EL, ET, CUE and CR. The form of these likelihoods is simple, being the product of the weights generated by the GEL procedures when the criterion function corresponds to EL, ET, CUE and CR, respectively.

Inference based on the semiparametric likelihoods will be sensitive to the assumptions under which the likelihoods are derived. A strong assumption is that all the possible values of the random variable have been observed. When this assumption is violated, the prior will be effectively data-dependent. In an idealized Bayesian view the prior should consist of information separated from the data and the model at hand. However, we point out that data- and model-dependence are common to the noninformative prior literature. We have addressed the issue of the validity of the resulting likelihoods and we find that the prior elicitation affects inference far more dramatically than the nonparametric assumptions. However, more research on this issue is needed.

It would also be extremely interesting to study the frequentist properties of the semiparametric likelihoods derived in this paper. By extending the work of Chernozhukov and Hong (2003), Ragusa (2006b) shows that normal asymptotic theory applies. He also shows that the estimators defined as the maximum of these likelihoods enjoy some interesting higher order properties, such as higher order efficiency after bias correction. An investigation of the coverage properties of the credible regions in repeated samples would be a welcome addition to the literature.

A Proofs of Propositions

Proof of Proposition 1

The proof makes heavy use of the results in Theorem 3.1 and Theorem 3.3 of Csiszar (1975). The functions in $b_x(\omega,\beta)$ are linearly independent modulo P_ω if $c'b_x(\omega,\beta) = 0$, $c = (c_1,\ldots,c_m) \in \mathbb{R}^m$, if and only if $c_1 = c_2 = \ldots = c_m = 0$ for all ω with $P_\omega(\omega) > 0$. The functions constituting $b_x(\omega,\beta)$ are linear combination of the columns of $G(\beta)$, $b_x(\omega,\beta) = (G_1(\beta)\omega,\ldots,G_m(\beta)\omega)$. where $G_k(\beta)$ denotes the kth column of $G(\beta)$. Linearly independence of the columns of $b_x(\omega,\beta)$ implies that $(c_1G_1(\beta) + \cdots + c_mG_m(\beta))\omega = 0$ if and only if $c_1 = c_2 = \ldots = c_m = 0$. By assumption, the column of $G(\beta)$ are linearly independent and this suffice to show linear independence of $b_x(\omega,\beta)$ if $P_\omega(\omega = 0) < 1$. By linearly independence of $b_x(\omega,\beta)$ and openness of $T_{P_\omega}(\beta)$ the interior of A_m is non-empty (see Csiszar (1975) remarks at page 156). Since by assumption $0 \in \bar{A}_{P_\omega}$, the I-projection exists and it is of form

$$P_{\omega\beta} = \begin{cases} c_x(\beta)\cdot\exp(t'b_x(\omega,\beta))P_\omega & \text{if } \omega \notin N \\ 0 & \text{if } \omega \in N \end{cases}$$

To determine $c_x(\beta)$ notice that $\int \exp(t'b_x(\omega,\beta))dP_\omega = 1$ implies the $c_x(\beta)$ given in the proposition and that integrability and openness of T_{P_ω} implies that the solution is feasible, $\int b_x(\omega,\beta)\exp(t'b_x(\omega,\beta))dP_\omega = 0$, for t that solves the strictly convex minimization $\min_{\tau\in T_{P_\omega}} \int \exp(t'b_x(\omega,\beta))dP_\omega$.

Proof of Proposition 2

The proof simply consists in rearranging the $P_{\omega\beta}$ and showing that the measure $P_{\omega\beta}$ can be rewritten as the product $P_{\omega\beta,1} \times \ldots \times P_{\omega\beta,n}$. Notice that

$$\begin{aligned}
& \exp\left(t'b_x(\omega,\beta) - \log c_x(\beta)\right) P_\omega \\
= \; & \exp\left(\sum_{i=1}^n \omega_i f_i(\beta,t) - \log\int \exp\left(\sum_{i=1}^n \omega_i f_i(\beta,t)\right) dP_\omega\right) P_\omega \\
= \; & \left\{\prod_{i=1}^n \exp\left(\omega_i f_i(\beta,t) - \sum_{i=1}^n \int \exp(\omega_i f_i(\beta,t))dP_{\omega,i}\right)\right\} P_{\omega,1}\times\ldots\times P_{\omega,n} \\
= \; & P_{\omega\beta,1}\times\ldots\times P_{\omega\beta,n}
\end{aligned}$$

By similar arguments,

$$\begin{aligned} t &= \arg\min_{t\in T_{P_\omega}} \int \exp(\sum_{i=1}^{n} \omega_i f_i(\beta,t)) dP_\omega \\ &= \arg\min_{t\in T_{P_\omega}} \sum_{i=1}^{n} \int \exp(\omega_i f_i(\beta,\tau)) dP_{\omega,i} \end{aligned}$$

as required.

□

Proof of Proposition 3

By independence of ω_i $(i = 1, \ldots, n)$ under $P_{\omega\beta}$ and by rearranging

$$\begin{aligned} \bar{L}(\mathrm{x}|\beta) &= \int \prod_{i=1}^{n} \omega_i d(P_{\omega\beta,1} \times \cdots \times P_{\omega\beta,n}) \\ &= \int \prod_{i=1}^{n} \omega_i \exp(\omega_i f(\beta,t) - k_x(f_i(\beta,t))) dP_{\omega\beta,1} \times \cdots \times dP_{\omega\beta,n} \\ &= \prod_{i=1}^{n} \int \omega_i \exp(\omega_i f(\beta,t) - k_x(f_i(\beta,t))) dP_{\omega\beta,i} \\ &= \prod_{i=1}^{n} \phi_i \end{aligned}$$

Since the ϕ_i are the probabilities on x they are normalized to 1, giving the first result. The second result follows from Proposition 2.

□

Proof of Proposition 4

For part (a), we need to show that the derivatives of the cumulant generating function of the Normal distribution with mean 1 and variance σ^2 corresponds to the likelihood objective functions in (5) and the cumulant generating function. The cumulant generating function of $N(1, \sigma^2)$ is given by $k_x(\upsilon) = \upsilon + \sigma^2\upsilon^2/2$ and hence the numerator of

$$k_{x,\partial}(\upsilon_i) = (1 + \upsilon_i)/\sum_{i=1}^{n}(1 + \upsilon_i).$$

For part (b), we need to show that the derivatives of the cumulant generating function of the Poisson distribution with mean 1 corresponds to the likelihood objective functions in (6). The cumulant generating function of $Poisson(1)$ is given by $k_x(\upsilon) = \exp(\upsilon) - 1$ and it follows that, in this case,

$$k_{x,\partial}(\upsilon_i) = \exp(\upsilon_i) / \sum_{i=1}^{n} \exp(\upsilon_i).$$

For part (c), we need to show that the derivatives of the cumulant generating function of the Gamma distribution with parameters 1 and ς corresponds to the likelihood objective functions in (7). The cumulant generating function of $Gamma(1, \varsigma)$ is given by $k_x(\upsilon) = -\log(1 - \varsigma\upsilon)$ and it follows that, in this case,

$$k_{x,\partial}(\upsilon_i) = (1 - \varsigma\upsilon_i)^{-1} / \sum_{i=1}^{n} (1 - \varsigma\upsilon_i)^{-1}.$$

Noting that in each case of the proposition the parameters τ is the $\arg\min_{t \in T(\beta)} \sum_{i=1}^{n} k_x(u_i(\beta, t))$, where $T(\beta) = \{t : \max_i u_i(\beta, t) < \infty\}$, gives the result.

□

Proof of Proposition 5

Let $\omega = \sum_{i=1}^{L} y_i$, where y_i $(i = 1, \ldots, L)$ are iid with density $p(y) = \frac{\delta^\delta}{\theta^\delta \Gamma(\delta)} e^{-\frac{\delta}{\theta} y} y^{\delta - 1}$, $(\delta > 0$, $\theta > 0$, and $y > 0)$, and $L \sim Poisson(\lambda)$. The moment generating function of $p(y)$

$$\begin{aligned}
\mathcal{L}_y(t) &\equiv \frac{\delta^\delta}{\theta^\delta \Gamma(\delta)} \int_0^{+\infty} e^{-\frac{\delta}{\theta} y} y^{\delta - 1} e^{ty} dy \\
&= \frac{\delta^\delta}{\theta^\delta \Gamma(\beta)} \int_0^{+\infty} e^{(t - \frac{\delta}{\theta}) y} y^{\delta - 1} dy \\
&= \frac{\delta^\delta}{\theta^\delta \Gamma(\delta)} \frac{\Gamma(\delta)}{(\frac{\theta}{\delta} - y)^\delta} \\
&= \frac{\delta^\delta}{\theta^\delta (\frac{\delta}{\theta} - y)^\delta} \\
&= \left(\frac{\delta}{\delta - \theta t} \right)^\delta, \quad \theta t < \delta
\end{aligned}$$

Since L and y_j $(j = 1, \ldots, L)$ are independent, the moment generating function of $\omega = \sum_{i=1}^{L} y_i$ is

$$\begin{aligned}
\mathcal{L}_\omega(t) &\equiv \sum_{k=0}^{\infty} \left[\int e^{t \sum_{j=1}^{L} y_j} p(y) dy \right] p(L = k) \\
&= \sum_{k=0}^{\infty} \mathcal{L}_y(t)^k \frac{\lambda^k e^{-\lambda}}{k!} \\
&= e^{\mathcal{L}_y(t)\lambda} e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda \mathcal{L}_y(t))^k e^{(\lambda \mathcal{L}(t))}}{k!} \\
&= e^{\mathcal{L}_y(t)\lambda} e^{-\lambda}
\end{aligned}$$

where the last equality holds because $(\lambda \mathcal{L}_y(t))^k e^{(\lambda \mathcal{L}(t))}/k!$ is the pdf of $Poisson(\lambda \mathcal{L}_y(t))$ and, hence, $\sum_{k=0}^{\infty} \frac{(\lambda \mathcal{L}_y(t))^k e^{(\lambda \mathcal{L}(t))}}{k!} = 1$. It follows that the logarithm of the moment generating function of the distribution of ω is given by $\log \mathcal{L}_\omega(t) = \lambda \mathcal{L}_y(t) - \lambda$. Thus, if each ω_i $(i = 1, \ldots, n)$ is $\omega_i = \sum_{j=1}^{L} y_j$, $y_j \sim p(y)$ and $L \sim Poisson(\lambda)$ the likelihood, after normalization, is given by

$$\prod_{i=1}^{n} (\varphi_i / \sum_{i=1}^{n} \varphi_i), \quad \varphi_i \equiv \left. \frac{\partial(\lambda \mathcal{L}_y(t) - \lambda)}{\partial t} \right|_{t = f_i(\beta, t)} = \lambda\theta \left(\frac{\delta}{\delta - \theta f_i(\beta, t)} \right)^{\delta + 1}$$

where

$$t = \arg \min_{\tau \in T_{\delta,\theta}(\beta)} \left[\lambda \sum_{i=1}^{n} \left(\frac{\delta}{\delta - \theta f_i(\beta, \tau)} \right)^{\delta} - \lambda \right], \quad T_{\delta,\theta}(\beta) = \{\tau : \delta - \theta f_i(\beta, \tau) > 0\}$$

It then follows that for $\gamma < -1$ and we have that $\varphi_i = \lambda(\gamma + 1)(1 + \gamma f_i(\beta, t))^{1/\gamma}$ where $\gamma = -\theta/\delta$ and $\delta = -\gamma/(\gamma + 1)$. Hence, the likelihood obtained is equivalent, up to proportionality constant, to the likelihood that one would obtain by multiplying the normalized weights of the GEL with a Cressie Read criterion, for $\gamma < -1$.

□

B Proof of Equation (11)

The parameter τ for EL, ET, CUE and CR based likelihood is implicitly defined by

$$\sum_{i=1}^{n} \rho_1(\tau g_i(\beta_q)) g_i(\beta_q) = 0,$$

where $\rho_1(u) = (1+u)^{-1}$ for EL, $\rho_1(u) = \exp(u)$ for ET, $\rho_1(u) = (1+u)$ for CUE, and $\rho_1^{CR}(u) = (1+\gamma u)^{1/\gamma}$ for CR. Since $g_i(\beta_q) = 1(x \leq \beta_q) - \tau$, the equation defining τ can be rewritten as

$$k(\beta_q)\left[\rho_1(\tau(1-q))(1-q)\right] + (n - k(\beta_q))\left[\rho_1(-\tau q)q\right] = 0$$

The likelihoods have the same basic structure:

$$\bar{L}(\mathrm{x}|\beta) \propto \left[\varphi_1(\beta_q)/r(\beta_q)\right]^{k(\beta_q)} \left[\varphi_2(\beta_q)/r(\beta_q)\right]^{n-k(\beta_q)}, \tag{13}$$

where

$$r(\beta_q) = \sum_{i=1}^{n} \rho_1(\tau g_i(\beta_q))$$

The proof for ET has been given by Lancaster and Jae Jun (2006). The other three cases are discussed separately below.

(a) For the EL the likelihood a closed-form solution for τ is obtained by solving

$$\sum_{i=1}^{n} \frac{\delta_i(\beta_q) - q}{1 + \tau(\delta_i(\beta_q) - q)} = 0, \quad \delta_i(\beta_q) = 1(x_i \leq \beta_q) - q$$

Solving gives

$$\tau = \frac{k(\beta_q)q - (n - k(\beta_q))(1-q)}{q(1-q)n} \tag{14}$$

notice that for EL, $r(\beta_q) = n$. Substituting 14 into the formula for the weights gives

$$\varphi_1(\beta_q) = \frac{(1-q)}{k(\beta_q)}, \quad \varphi_2(\beta_q) = \frac{q}{(n - k(\beta_q))},$$

and hence the likelihood is proportional to

$$\left[\frac{(1-q)}{k(\beta_q)}\right]^{k(\beta_q)}\left[\frac{q}{n-k(\beta_q)}\right]^{n-k(\beta_q)} \propto \left[\frac{q}{k(\beta_q)}\right]^{k(\beta_q)}\left[\frac{1-q}{n-k(\beta_q)}\right]^{n-k(\beta_q)}$$

where the last proportionality follows from multiplying the likelihood by $(q/1-q) = (q/1-q)^{k(\beta_q)}(q/1-q)^{n-k(\beta_q)}$ and simplifying.

(b) For the CUE we have that $\sum_{i=1}^{n}(1+\tau g_i(\beta_q)) = n+\tau(k(\beta_q)-nq)$. Tedious but straightforward algebraic manipulations show that

$$\tau = -\left(k(\beta_q)(1-q)^2 + (n-k(\beta_q))q^2\right)^{-1}(k(\beta_q)-nq) \tag{15}$$

The sum of the weights of the CUE is given by

$$r(\beta_q) = \frac{k(\beta_q)(n-k(\beta_q))}{k(\beta_q)(1-q)^2+(n-k(\beta_q))q^2}$$

Using the above formula for $r(\beta_q)$, substituting 15 into the formula for the weights and simplifying gives

$$\varphi_1(\beta_q)/r(\beta_q) = \frac{q}{k(\beta_q)}, \quad \varphi_2(\beta_q)/r(\beta_q) = \frac{1-q}{n-k(\beta_q)}$$

Substituting the above expressions into (13) proves the result for the CUE-based likelihood.

(c) For the CR case τ is given, for $\gamma < -1$, by

$$\tau = \frac{1-c}{c\gamma(1-q)+\gamma q}, \quad c = \left[\frac{k(\beta_q)(1-q)}{(n-k(\beta_q))q}\right]^{\gamma}$$

Simple, yet tedious calculations, reveal that

$$r(\beta_q) = k(\beta_q)\left[\frac{1}{c(1-q)+q}\right]^{1/\gamma} + (n-k(\beta_q))\left[\frac{c}{c(1-q)+q}\right]^{1/\gamma}$$

and hence

$$\varphi_1(\beta_q)/r(\beta_q) = \frac{q}{k(\beta_q)}, \quad \varphi_2(\beta_q)/r(\beta_q) = \frac{1-q}{n-k(\beta_q)}$$

Substituting the above expressions into (13) proves the result for the CR-based likelihood.

□

References

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*, New York, NY: Springer Verlag.

Brown, B. W. and W. K. Newey (2002): "Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference," *Journal of Business and Economic Statistics*, 20, 507–517.

Chamberlain, G. and G. W. Imbens (2003): "Nonparametric Applications of Bayesian Inference," *Journal of Business and Economic Statistics*, 21, 12–18.

Chernozhukov, V. and H. Hong (2003): "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293–346.

Csiszar, I. (1975): "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, 3, 146–58.

Ferguson, T. (1973): "A Bayesian Analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

——— (1974): "Prior of distributions on spaces of probability measures," *The Annals of Statistics*, 2, 615–629.

Gallant, R. A. and H. White (1988): *A unified theory of estimation and inference for nonlinear dynamic models*, New York: Basil Blackwell.

Gasparini, M. (1995): "Exact Multivariate Bayesian Bootstrap Distributions of Moments," *The Annals of Statistics*, 23, 762–768.

Hahn, J. (1997): "Bayesian Bootstrap of the Quantile Regression Estimator: A Large Sample Study," *International Economic Review*, 38, 206–213.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–54.

Imbens, G. W. (1997): "One-Step Estimators for Over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359–83.

Jaynes, E. T. (1968): "Prior Probabilities," *IEEE Transactions on System Science and Cybernetics*, SSC-4, 227–241.

Jeffreys, H. (1967): *Theory of Probability*, Oxford University Press, 3rd ed.

Kim, J. Y. (2002): "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, 107, 175–93.

Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments estimation," *Econometrica*, 65, 861–74.

Lancaster, T. (1994): "Bayes WESML: Posterior Inference from Choice-Based Samples," Unpublished manuscript, Brown University, Providence, RI.

Lancaster, T. and S. Jae Jun (2006): "Bayesian quantile regression," CeMMAP working papers CWP05/06, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, available at http://ideas.repec.org/p/ifs/cemmap/05-06.html.

Lazar, N. A. (2003): "Bayesian empirical likelihood," *Biometrika*, 90, 319–326.

Monahan, J. F. and D. Boos (1992): "Proper likelihoods for Bayesian analysis," *Biometrika*, 79, 271–278.

Newey, W. K. and D. McFadden (1994): "Estimation and inference in large samples," in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, Amsterdam: North-Holland, 2113–2245.

Newey, W. K. and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–55.

Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *Annals of Statistics*, 22, 300–325.

Ragusa, G. (2006a): "Frequentist properties of a class of semiparametric Bayesian procedures," Tech. rep., University of California, Irvine.

——— (2006b): "Minimum Divergence, Generalized Empirical Lielihoods and Higher Order Expansions," Tech. rep., University of California, Irvine.

Rubin, D. (1981): "Bayesian Bootstrap," *Annals of Statistics*, 9, 130–34.

Schennach, S. C. (2005): "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika*, 92, 31–46.

Sims, C. A. (2002): "Interview with Christopher A. Sims," *Journal of Business and Economic Statistics*, 20, 448–49.

Figure 1: Comparison of posterior distributions.

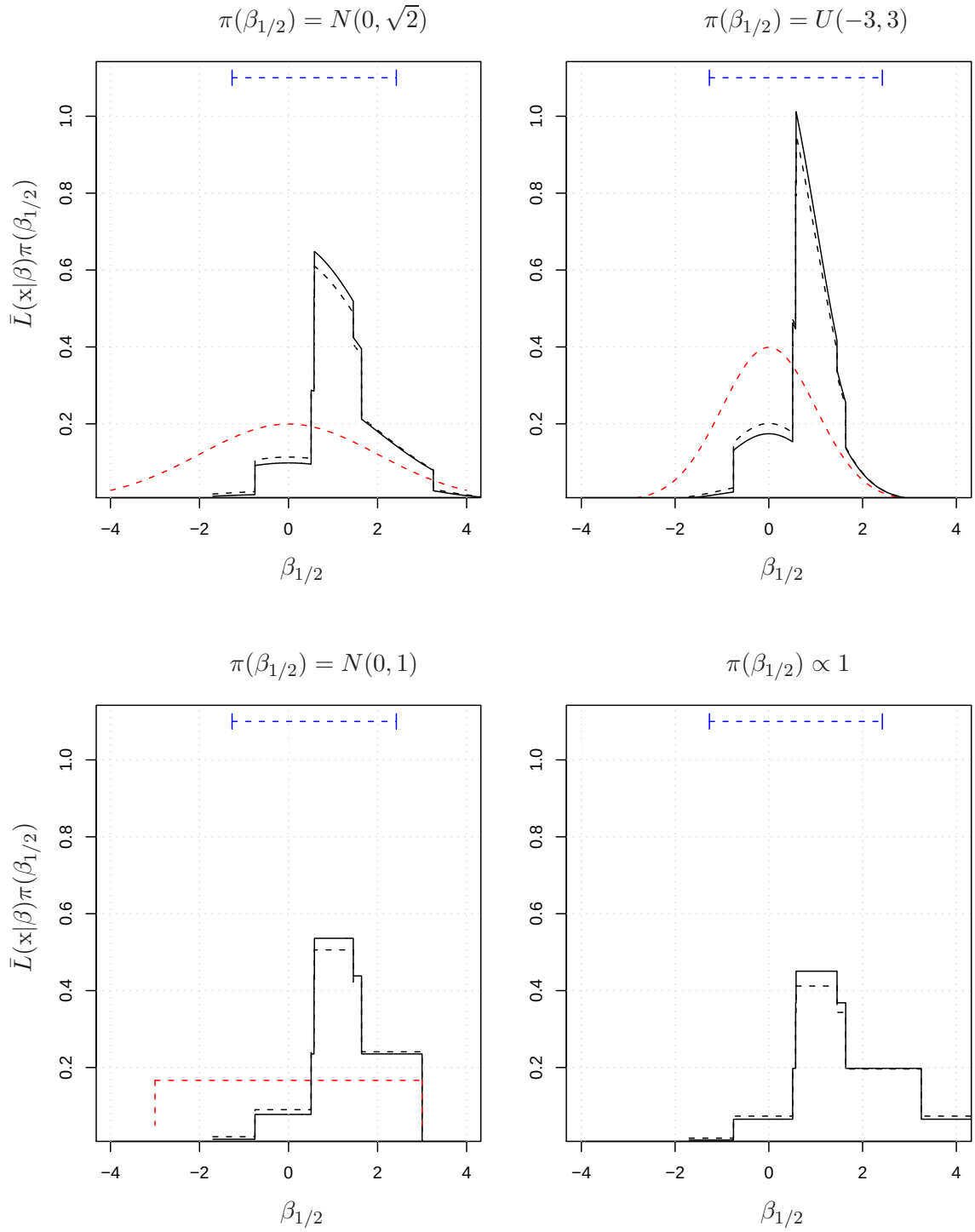

Figure 2: Quantiles to quantiles plot of the $a(\beta)$ criterion for different values of the variance of the prior distribution of β.

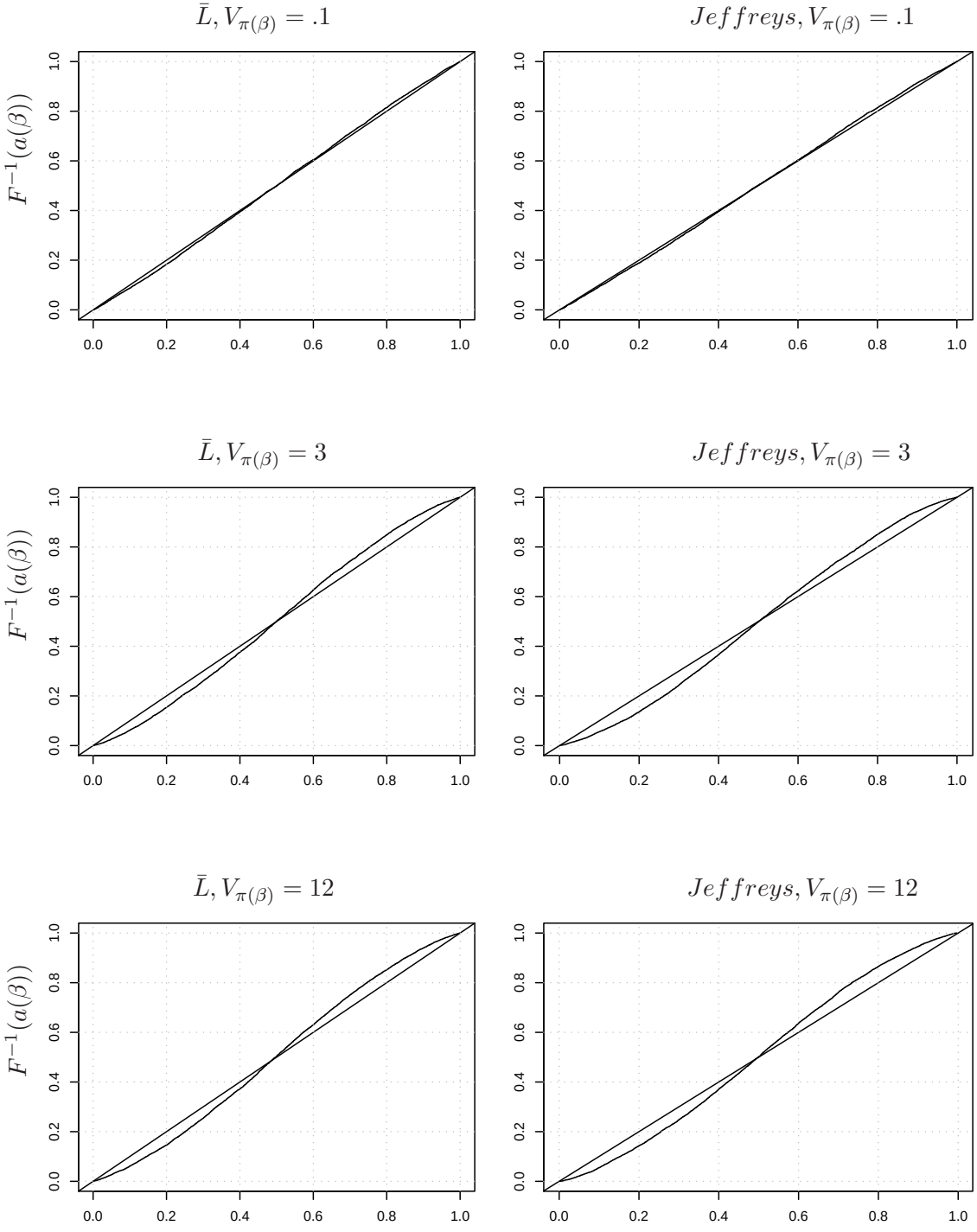

Figure 3: Histograms of the $a(\beta)$ criterion for different values of the variance of the prior distribution of β.

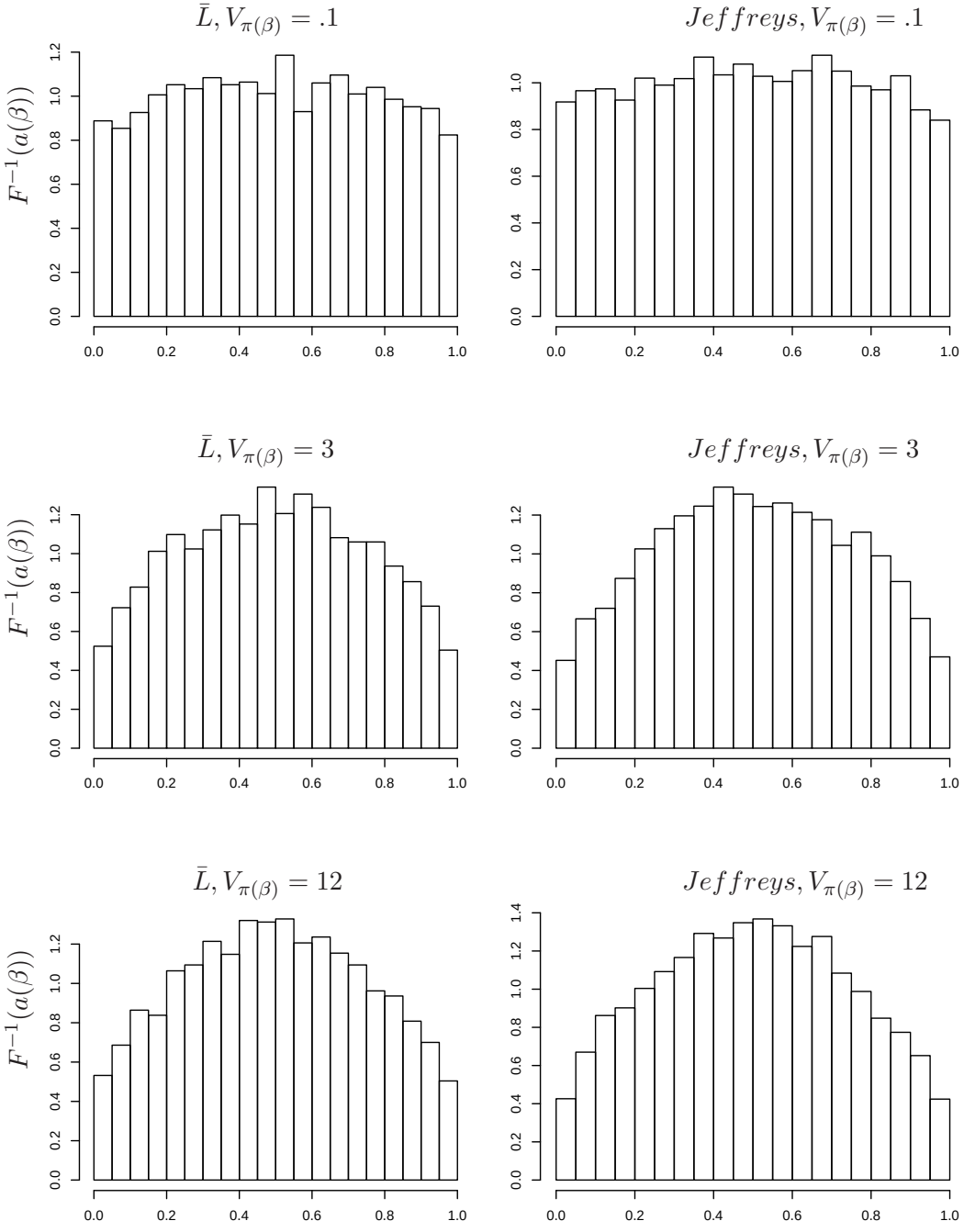

Figure 4: pvalues of the KS statistics for testing that $a(\beta) \sim U(0,1)$. pvalues are plotted for each sample size and prior considered.

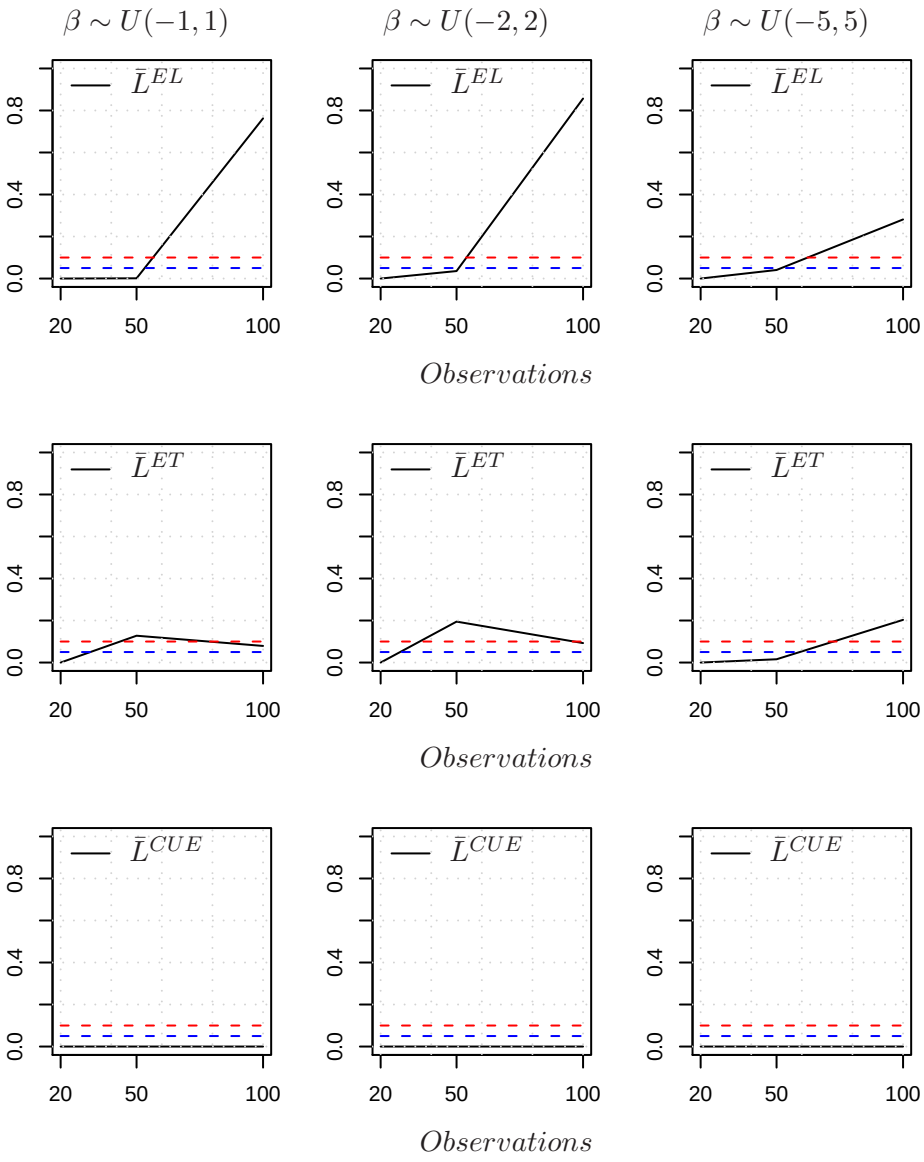

Figure 5: Quantiles to Quantile plot of the $a(\beta)$ criterion: $\beta = U(-1, 1)$

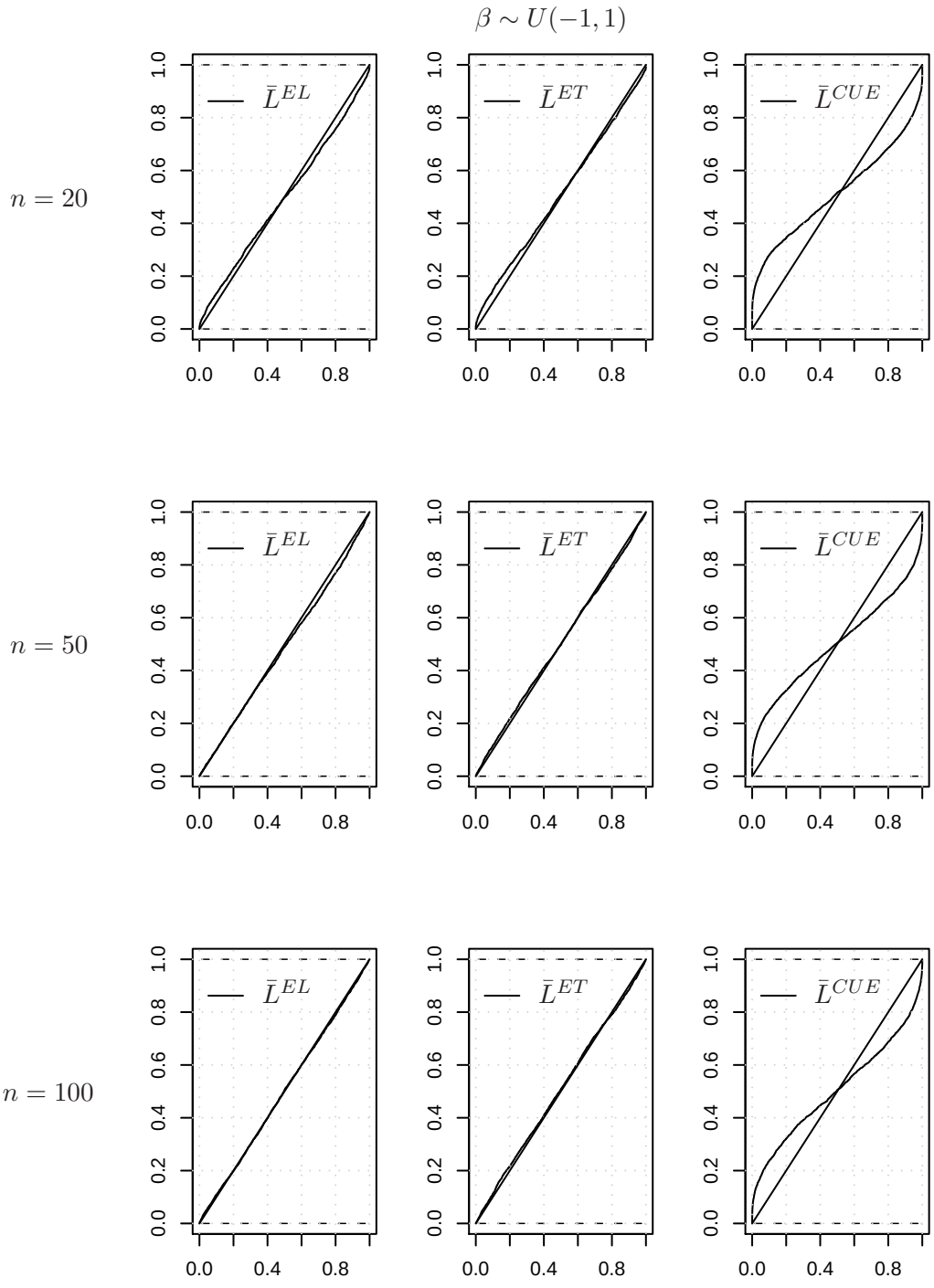

Figure 6: Quantiles to Quantile plot of the $a(\beta)$ criterion: $\beta = U(-2, 2)$

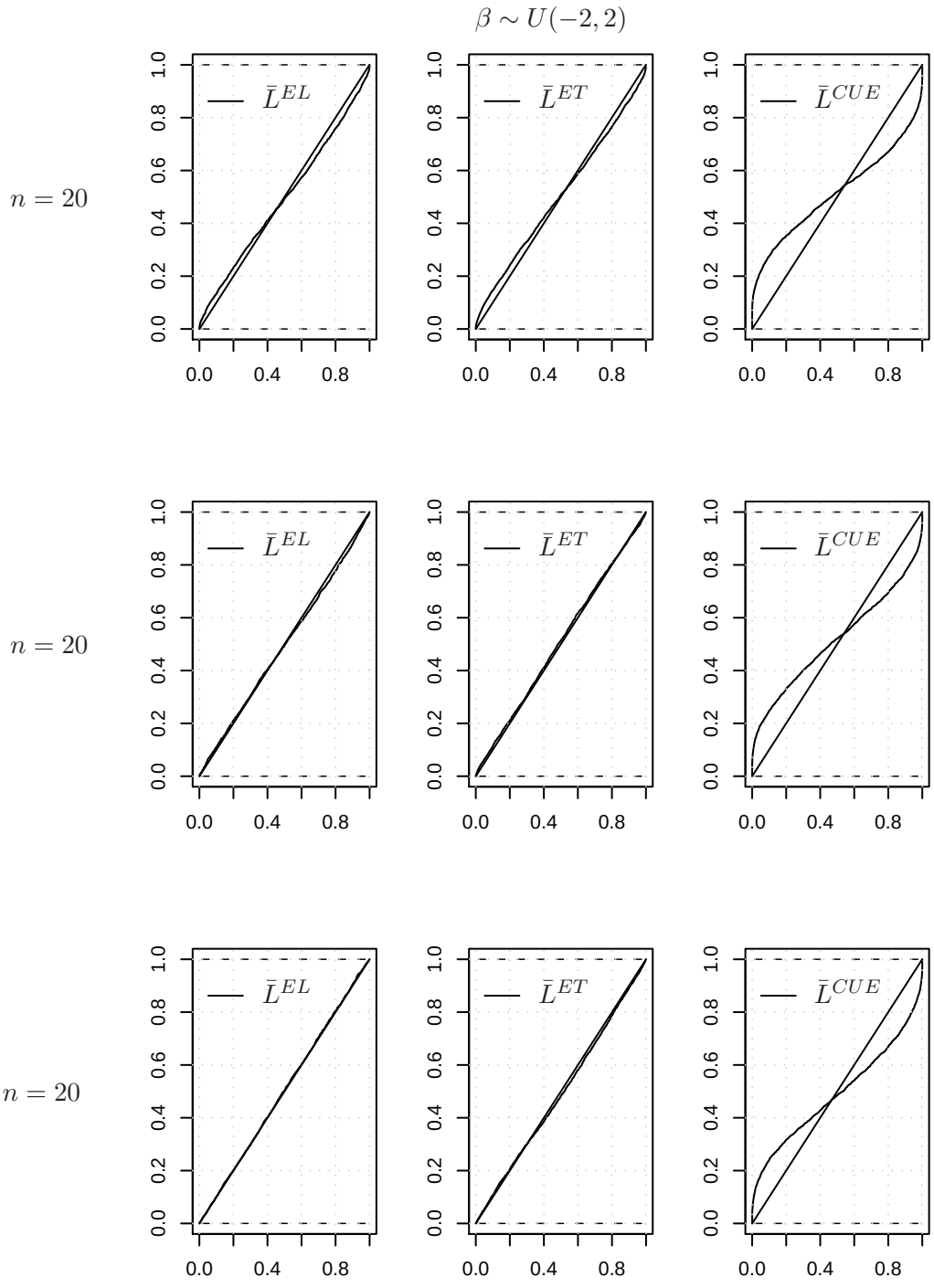

Figure 7: Quantiles to Quantile plot of the $a(\beta)$ criterion: $\beta = U(-5, 5)$